DATA PAPER

Medical Named Entity Recognition from Un-labelled Medical Records based on Pre-trained Language Models and Domain Dictionary

Chaojie Wen, Tao Chen†, Xudong Jia & Jiang Zhu

Faculty of Intelligent Manufacturing, Wuyi University, Jiangmen 529020, China

Keywords: Medical named entity recognition; Pre-trained language model; Domain dictionary; Pseudo labelling; Un-labelled medical data

Citation: Wen, C.J., et al.: Medical named entity recognition from un-labelled medical records based on pre-trained language models and domain dictionary. Data Intelligence 3(3), 402-417 (2021). doi: 10.1162/dint_a_00105
Received: February 15, 2021; Revised: March 29, 2021; Accepted: April 22, 2021

ABSTRACT

Medical named entity recognition (NER) is an area in which medical named entities are recognized from medical texts, such as diseases, drugs, surgery reports, anatomical parts, and examination documents. Conventional medical NER methods do not make full use of un-labelled medical texts embedded in medical documents. To address this issue, we proposed a medical NER approach based on pre-trained language models and a domain dictionary. First, we constructed a medical entity dictionary by extracting medical entities from labelled medical texts and collecting medical entities from other resources, such as the Yidu-N4K data set. Second, we employed this dictionary to train domain-specific pre-trained language models using un-labelled medical texts. Third, we employed a pseudo labelling mechanism in un-labelled medical texts to automatically annotate texts and create pseudo labels. Fourth, the BiLSTM-CRF sequence tagging model was used to fine-tune the pre-trained language models. Our experiments on the un-labelled medical texts, which were extracted from Chinese electronic medical records, show that the proposed NER approach enables the strict and relaxed *F*1 scores to be 88.7% and 95.3%, respectively.

† Corresponding author: Tao Chen (Email: chentao1999@gmail.com; ORCID: 0000-0002-3634-0854).

1. INTRODUCTION

Medical records are important resources in which patients' diagnosis and treatment activities in hospitals are documented. In recent years, many medical institutions have done significant work in archiving electronic medical records. Handwritten medical records are gradually being replaced by digital ones. Many researchers strive for extracting medical knowledge from digital data, using medical knowledge to help medical professionals understand potential causes of various symptoms, and building medical decision support systems.

Medical named entity recognition (NER) is an important technique that has recently received attention in medical communities in extracting named entities from medical texts, such as diseases, drugs, surgery reports, anatomical parts, and examination documents [1]. For example, researchers within the Chinese Information Processing Society of China (CIPSC) have established a program entitled the CCKS 2020 Medical Entity and Event Extraction from Chinese Electronic Medical Records (or Chinese EMRs). This program aims at identifying medical entities from archived Chinese EMRs which are in plain text format. Additionally, the program groups medical entities into six pre-defined categories: disease & diagnosis, imaging examination, laboratory examination, surgery, drug, and anatomic part (CCKS 2020 Medical Entity and Event Extraction from Chinese EMRs Task 1: Medical Entity Recognition)[①]. As an important contributor to this program, Yidu Cloud Co., Ltd. has recently built the first medical data set from the Chinese electronic medical records (this data set is referenced later to as the CCKS-2020-CEMR data set.) This data set contains a dictionary of 6,292 words, 1,000 un-labelled texts, and 1,500 labelled texts. In the labelled texts, 26,414 medical entities are annotated. It is anticipated that this data set will be evolved to be the benchmark data set for medical named entity recognition. We used this data set to test and evaluate our NER approach.

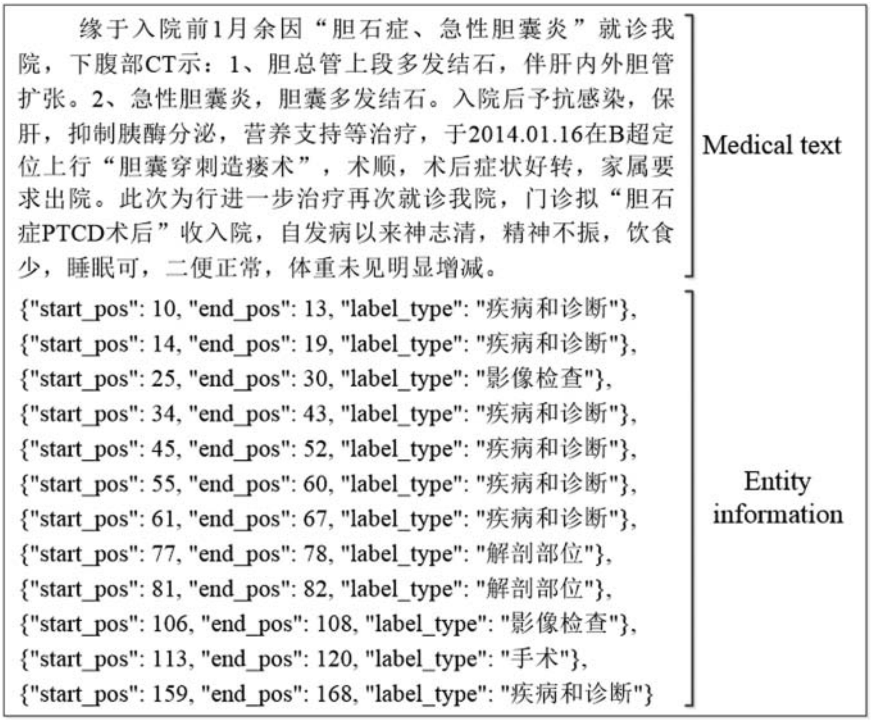
The CCKS-2020-CEMR data set is summarized in Table 1. An example of labelled texts is shown in Figure 1.

Table 1. Summary of the CCKS-2020-CEMR data set.

#Text	#Disease & diagnosis	#Imaging examination	#Laboratory examination	#Surgery	#Drug	#Anatomic part	Total
1,500	6,211	1,490	1,885	1,327	2,841	12,660	26,414

There are medical NER methods and models developed for extracting medical knowledge from data sets. However, these methods and models are supervised methods which help us learn features from labelled data. In these methods, un-labelled data cannot be used effectively and are often discarded. As a result, medical information hidden in un-labelled texts cannot be recognized. To address this issue, we proposed a medical NER approach based on the combination of pre-trained language models and a domain dictionary. We employed the pre-trained language models and the in-domain training technique to extract medical knowledge from un-labelled medical texts.

[①] https://www.biendata.xyz/competition/ccks_2020_2_1



缘于入院前1月余因“胆石症、急性胆囊炎”就诊我院，下腹部CT示：1、胆总管上段多发结石，伴肝内外胆管扩张。2、急性胆囊炎，胆囊多发结石。入院后予抗感染，保肝，抑制胰酶分泌，营养支持等治疗，于2014.01.16在B超定位上行“胆囊穿刺造瘘术”，术顺，术后症状好转，家属要求出院。此次为行进一步治疗再次就诊我院，门诊拟“胆石症PTCD术后”收入院，自发病以来神志清，精神不振，饮食少，睡眠可，二便正常，体重未见明显增减。

```
{"start_pos": 10, "end_pos": 13, "label_type": "疾病和诊断"},
{"start_pos": 14, "end_pos": 19, "label_type": "疾病和诊断"},
{"start_pos": 25, "end_pos": 30, "label_type": "影像检查"},
{"start_pos": 34, "end_pos": 43, "label_type": "疾病和诊断"},
{"start_pos": 45, "end_pos": 52, "label_type": "疾病和诊断"},
{"start_pos": 55, "end_pos": 60, "label_type": "疾病和诊断"},
{"start_pos": 61, "end_pos": 67, "label_type": "疾病和诊断"},
{"start_pos": 77, "end_pos": 78, "label_type": "解剖部位"},
{"start_pos": 81, "end_pos": 82, "label_type": "解剖部位"},
{"start_pos": 106, "end_pos": 108, "label_type": "影像检查"},
{"start_pos": 113, "end_pos": 120, "label_type": "手术"},
{"start_pos": 159, "end_pos": 168, "label_type": "疾病和诊断"}
```

Figure 1. An example of the CCKS-2020-CEMR data set, where the “start_pos”, “end_pos”, and “label_type” terms refer to the start position, end position, and medical label of the entity, respectively.

The main contributions of our work are summarized below:

1) We constructed a medical domain entity dictionary and integrated it into pre-trained language models so that the recognition rate (measured by $F1$ score) of named entities has been significantly improved.
2) We fed un-labelled medical texts to pre-trained language models for in-domain training. Experimental results show that the proposed NER approach with the help of the in-domain training technique can improve the performance of named entity recognition in medical communities.
3) A pseudo labelling method was used to augment the CCKS-2020-CEMR data set by automatically annotating un-labelled texts and marking them as pseudo labels.

The rest of this paper is organized as follows: Section 2 describes previous research work relative to medical NER. With a good understanding of the state-of-the-art practice in medical knowledge extraction, we present our medical NER approach in Section 3. Section 4 describes the evaluation of the medical NER approach when this approach is used for recognition of medical knowledge from the CCKS-2020-CEMR data set. The evaluation results of the NER approach with the help of the pre-trained language models and an entity dictionary are also analyzed in this section. The paper is concluded in Section 5 by summarizing the contributions of the research work and outlining the future research directions.

2. PREVIOUS RESEARCH WORK ON MEDICAL NER

Medical NER [2] is a basic task for extracting knowledge representation from electronic medical records. Solutions to medical NER can be roughly categorized into two categories: traditional machine learning based methods and neural network based methods. The traditional machine learning based methods mainly treat NER as a sequence labelling task. Wang et al. [3] conducted a study by using conditional random fields (CRFs), support vector machines (SVM), and maximum entropy (ME) to recognize symptoms and pathogenesis in Chinese EMRs. Wang et al. [4] investigated the use of CRF in different feature sets and recognized symptom names from clinical notes of traditional Chinese medicine. Xu et al. [5] proposed a joint model that integrated segmentation and NER simultaneously to improve the performance of NER in Chinese discharge summaries. These methods rely heavily on hand-crafted features.

Unlike the conventional machine learning based methods, the neural network based methods have a training process (or an end-to-end process) which automatically learns from data features. There are a group of studies dealing with medical NER using neural network based methods. Among these studies, Wu et al. [6] developed a deep neural network approach for NER in Chinese clinical texts. Yang et al. [7] combined the characteristics of Chinese electronic language structure, developed labelling rules for named entity recognition in Chinese electronic medical records, and completed the natural language processing research in extracting knowledge from Chinese EMRs. Yang et al. [8] used the combination model of Bi-directional Long Short-Term Memory (BiLSTM) and CRF to extract medical entities from admission records and discharge summaries. Additionally, Chowdhury et al. [9] proposed a multitask bi-directional Recurrent Neural Network (RNN) model to recognize diseases, symptoms, and other entities in Chinese electronic medical records. The RNN model employed two different task layers (the word embedding and character embedding layers) to improve the accuracy of extracting entity information from EMRs data set. Furthermore, Wan et al. [10] proposed a Chinese medical NER method based on joint training of Chinese characters and words. This model took Chinese language features into account and added semantic information of words in its NER process.

All these medical NER methods and models do not make full use of un-labelled medical texts embedded in medical documents. Therefore, developing a medical NER approach that can recognize named entities from un-labelled texts is very important.

3. NER METHODOLOGY FOR MEDICAL TEXTS

In this section, we present our approach for medical NER using pre-trained language models and a domain dictionary. An overview of the approach is shown in Figure 2. First, we preprocessed labelled medical texts (in Section 3.1) and constructed a medical entity dictionary (in Section 3.2). Second, we introduced a pseudo labelling method and used this method to annotate un-labelled texts with pseudo labels (in Section 3.3). Third, we used the BiLSTM-CRF technique to fine-tune the pre-trained language models (in Section 3.4). Fourth, we employed medical domain pre-trained language models (including the WoBERT model) and trained them using un-labelled medical texts (in Section 3.5). In the end, we combined the results from different pre-trained models for named entity recognition (in Section 3.6).

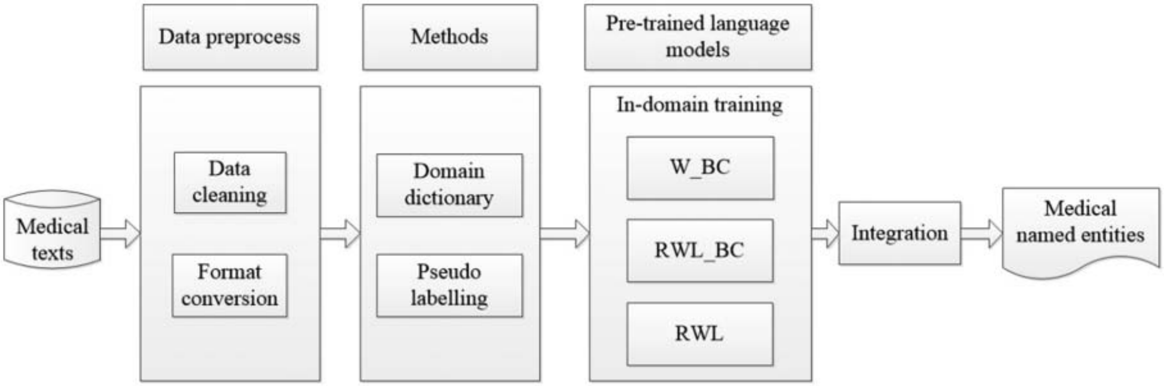


Figure 2. Framework of our approach. The W_BC, RWL_BC, and RWL terms refer to the WoBERT_BiLSTM-CRF model, the RoBERTa-wwm-ext-large_BiLSTM-CRF model and the RoBERTa-wwm-ext-large model, respectively. These three models will be described later.

3.1 Data Preprocess

Before we implemented our medical NER approach to extract knowledge from the labelled texts (embedded in the CCKS-2020-CEMR data set), we found that the CCKS-2020-CEMR data set has errors and needs to be preprocessed. The data preprocess task consists of the following four steps:

1) Data cleaning. In this step, we removed obvious labelling errors in the data set. For example, we removed extra spaces in “直肠癌根　治术” and “胃角　高级别下皮内瘤变” (In English, “radical resection of rectal cancer” and “gastric angle with high grade endothelial neoplasia”). We also fixed entity boundary errors.
2) Text normalization. In this step, we converted full angle Chinese symbols to half angle symbols. In addition, we changed uppercase English letters to lower case letters.
3) Corpus format conversion. In this step, we converted labelled texts from CEMR format to BIO format. In the BIO format, each element is labelled as “B-X”, “I-X” or “O”, where “B-X” refers to the beginning of an entity with type X (Type X refers to one of the six medical categories: disease & diagnosis, imaging examination, laboratory examination, surgery, drug, and anatomic part), “I-X” refers to the middle of an entity with Type X, and “O” indicates that the word is not a medical entity. The 1,500 BIO formatted texts were used as the training data for our proposed medical NER approach.
4) Corpus segmentation. Considering the limitation of the LSTM method in dealing with long sequences, we, in this study, segmented medical records into shorter sequences while we ensured the relative integrity of sentences in the records. Each sequence was limited to 200 characters. Period and other symbols were used to retain semantic information. [CLS] and [SEP] tokens were also added.

3.2 Domain Dictionary

There is a medical dictionary containing 6,292 medical words in the CCKS-2020-CEMR data set. To expand this dictionary, we have crawled related medical words from public medical websites[②]. Duplicate or highly similar medical words were removed. At end, we collected 11,000 medical words which were related to drugs, diseases, anatomic parts, and other types of medical information.

We used the word-piece [11] format to represent the medical words in our medical dictionary. The word-piece format is a kind of text representation which divides English words into pieces. First, we divided words into sub-words in the given training set. Then we added special word boundary symbols before we tested and evaluated our NER approach. In doing so, original word sequences, while kept unchanged, can be recovered from text sequences. For example, the word-piece representation of "the highest mountain" is "the high ##est moun ##tain", in which the word "highest" is divided into two pieces: "high", and "##est", while the word "mountain" is divided into "moun", and "##tain".

The word-piece format is also suitable for Chinese word representation. For example, a Chinese medical entity "左侧颈动脉中段" (mid left carotid artery) is split into four pieces: "左侧" (left), "## 颈" (neck), "## 动脉" (artery), "## 中段" (middle). Using the word-piece format, we segmented Chinese words from the 11,000 medical entities (stored in our medical dictionary), converted them into 30,000 word pieces, and put all these word pieces into the original dictionaries in pre-trained language models.

3.3 Pseudo Labelling

During the time when we used the WoBERT_FPT model for the in-domain training process, we also adopted a pseudo labelling method and augmented labelled texts for this study. Pseudo labelling [12] is an unsupervised learning method that improves the performance of an NER model in extracting knowledge from un-labelled texts. The purpose of our pseudo labelling method is to use labelled texts as a training set to train our BiLSTM-CRF model first, and then use the model to predict labels for un-labelled texts. The predicted labels are named as pseudo labels. Labelled texts and pseudo-labelled texts are further combined as a new training set for named entity recognition.

The framework for the pseudo labelling method is shown in Figure 3. Labelled texts are used to train the RoBERTa-wwm-ext-large-BiLSTM-CRF model. The trained RoBERTa-wwm-ext-large-BiLSTM-CRF model then predicts pseudo-labels from the un-labelled texts. The pseudo-labelled texts with a high confidence level are added into the original training set.

[②] Yidu-N7K data set (data set of CHIP 2019 clinical terminology standardization task): http://openkg.cn/dataset/yidu-n7k

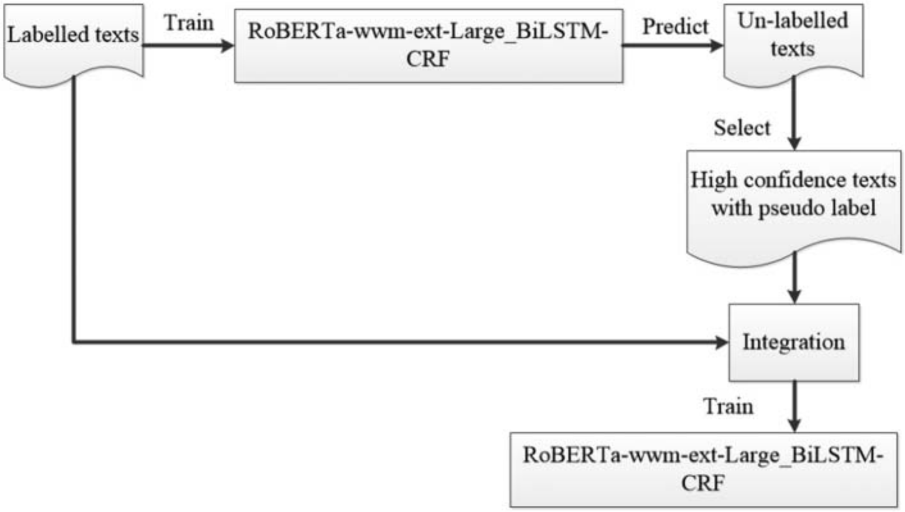

Labelled texts
Train
RoBERTa-wwm-ext-Large_BiLSTM-CRF
Predict
Un-labelled texts
Select
High confidence texts with pseudo label
Integration
Train
RoBERTa-wwm-ext-Large_BiLSTM-CRF

Figure 3. The framework of the pseudo labelling method.

3.4 Pre-trained Language Models

Three pre-trained models, namely "WoBERT_FurtherPreTraining_BiLSTM-CRF" (W_FPT_BC), "RoBERTa-wwm-ext-large" (RWL), and "RoBERTa-wwm-ext-large_BiLSTM-CRF" (RWL_BC), are developed in our NER approach to deal with un-labelled medical texts. All these three models use labelled and pseudo-labelled texts (described in the previous sections) as the training set and consider our medical dictionary to be the dictionary file.

It is noted that all these pre-trained models are derived from the RoBERTa model. The RoBERTa model [13] is a pioneer pre-trained language model whose focus is to train itself from a large-scale text corpus in an unsupervised way. It uses Chinese characters as the basic processing unit. The RoBERTa-wwm-ext-large model improves the RoBERTa model by implementing the Whole Word Masking (wwm) technique and masking Chinese characters that make up same words [14]. In other words, the RoBERTa-wwm-ext-large model uses Chinese words as the basic processing unit.

Pre-trained language models expand their own dictionary by adding the words in the domain dictionary (we have built and described in Section 3.2). When texts in the training set are mapped to those in the expanded dictionary, they are assigned to an identifier (or ID). Different words are mapped with different IDs. All out-of-vocabulary (OOV) words in the training set cannot be mapped with any words in the expanded dictionary and then are assigned with a unique ID. For example, a word "维生素D" (vitamin D) is not in the original dictionary of a pre-trained language model; however, it is inside the domain dictionary. Before the domain dictionary is merged into the original dictionary, the "维生素D" (vitamin D) word in our training data set is considered an OOV word. It cannot be mapped to that in the original dictionary, and thus is assigned with an ID of "100". After the domain dictionary is embedded in the original dictionary, the “维生素D” (vitamin D) word can be found in the expanded dictionary. It thus is assigned to an ID of “7818”.

Building on these three pre-trained models, our medical NER approach uses conventional sequence tagging models such as Bi-LSTM and CRF to fine-tune the pre-trained language models.

3.5 In-domain Training

Applying the pre-trained language models into Chinese medical NER directly may lead to unsatisfactory results due to a word distribution shift from general domain texts to Chinese medical texts. One way to alleviate this issue is to obtain domain-specific (or medical) knowledge through in-domain training. In-domain training, also known as continuous training [15], is a technique that uses pre-trained language models to extract domain-specific knowledge from texts. It can provide substantial gains over models that are trained on general knowledge or mixed domain knowledge.

In this paper, we used the WoBERT model for in-domain training. WoBERT[3] is a pre-trained language model derived from the RoBERTa-wwm-ext model. It takes advantage of the RoBERTa-wwm-ext model and optimizes its pre-training process when it works with un-labelled Chinese texts. The medical domain dictionary described in Section 3.2 is used as the training dictionary for the WoBERT model. The un-labelled texts in the CCKS-2020-CEMR data set are used as the training set for the NER approach. After training, a domain special language model is obtained. This model is further referred to as the WoBERT with Further Pre-Training model (WoBERT_FPT in short). The WoBERT_FPT model can be adapted for Chinese medical texts.

3.6 NER Models for Medical Texts

In this study we have developed nine NER models: 1) BiLSTM-CRF, 2) BERT_BC, 3) W_BC, 4) W_FPT_BC, 5) W_FPT_BC+pseudo labelling, 6) W_FPT_BC+domain dictionary+pseudo labelling, 7) RWL_BC+domain dictionary+pseudo labelling, 8) RWL+domain dictionary+pseudo labelling, and 9) fusion model.

The BiLSTM-CRF model does not have a pre-trained module before its sequence tagging process is executed. This model is a traditional model. It cannot extract hidden knowledge from un-labelled medical texts.

The BERT_BC model uses BERT as its pre-trained module for its named entity recognition. The W_BC model employs the WoBERT model as its in-domain pre-trained module before the BiLSTM-CRF (BC) sequence tagging process is conducted. The W_FPT_BC model also adopts the WoBERT model and expands its pre-training capabilities before the BC-based sequence tagging process is applied. The W_FPT_BC+pseudo labelling model adds pseudo labelled texts into the training set before the W_FPT_BC model is executed. This is our first NER model in dealing with un-labelled medical texts. The W_FPT_BC+domain

[3] WoBERT: https://github.com/ZhuiyiTechnology/WoBERT

dictionary+pseudo labelling model incorporates the domain dictionary and the pseudo labels in handling un-labelled medical texts. This is our second NER model for extracting knowledge from un-labelled medical texts.

The RWL_BC+domain dictionary+pseudo labelling model uses the "RoBERTa-wwm-ext-large" (RWL) model as its pre-trained module and works with the domain dictionary and the pseudo labels to recognize named medical entities through the BC method. This is our third NER model dealing with un-labelled medical texts. The RWL+domain dictionary+pseudo labelling model uses the domain dictionary and the pseudo labels in the RWL process. This is our fourth NER model. The fusion model combines the results from our NER models (that is, "WoBERT_FurtherPreTraining_BiLSTM-CRF" (W_FPT_BC), "RoBERTa-wwm-ext-large" (RWL), and "RoBERTa-wwm-ext-large_BiLSTM-CRF" (RWL_BC)) dealing with un-labelled medical texts.

4. EVALUATION

Precision (*P*), recall (*R*), and *F*1 score are computed as follows to evaluate the performance of the nine medical NER models.

$$P = \frac{|\mathrm{S} \cap \mathrm{G}|}{|\mathrm{S}|} \tag{1}$$

$$R = \frac{|\mathrm{S} \cap \mathrm{G}|}{|\mathrm{G}|} \tag{2}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{3}$$

where S refers to the predicted result set of the model, and G refers to the golden result set.

We have adopted two indicators, that is, strict and relaxed *F*1 scores, in this research to assess the effectiveness of the nine medical NER models in extracting medical knowledge. According to the boundary definition of recognized results, the strict indicator requires the predicted entity exactly equal to the golden result, while the relaxed indicator only requires the boundary of the predicted entity to be covered by the boundary of the golden result.

We have used the Adam Optimizer as our optimization algorithm to descend stochastic gradient for training our nine NER models. The Adam Optimizer combines the best properties of the AdaGrad and RMSProp algorithms and provides an optimization algorithm that can handle sparse gradients on noisy problems. In this research, the learning rate is set to be 5e-5, the dropout rate to be 0.5, and the batch size to be 5. The BiLSTM model is supported with 128 hidden layers. The strict and relaxed *F*1 scores of each model are shown in Tables 2 and 3, respectively.

It is noted from Tables 2 and 3 that using the pre-trained models, we can improve the effectiveness of the traditional NER models. For example, when the BERT pre-trained model is used before the BiLSTM-CRF model, the overall *F*1 score of the BiLSTM-CRF model increases by 6.9% (strict indicator) and 5.1% (relaxed indicator), respectively. Using the optimized pre-trained model, we also improved the accuracy of recognition. For example, the overall *F*1 score of the NER model combined with the WoBERT model is 2.2% (strict indicator) and 3.8% (relaxed indicator) higher than the NER model combined with the BERT model. After using in-domain training, we increased the overall *F*1 score of the WoBERT_BiLSTM-CRF model by 0.8% (strict indicator) and 0.9% (relaxed indicator), respectively. After the domain dictionary is added, the overall *F*1 score of the WoBERT_BiLSTM-CRF model is improved by 0.2% (strict indicator) and 0.5% (relaxed indicator), respectively. This indicates that using optimized pre-trained models, in-domain training technology, and the domain dictionary, we can improve the performance of the traditional medical NER models (or the BiLSTM-CRF and BERT_BC models) significantly.

Table 2. NER model results (strict).

	Disease & diagnosis	Imaging examination	Laboratory examination	Surgery	Drug	Anatomic part	Total
BiLSTM-CRF	0.660	0.787	0.711	0.709	0.862	0.807	0.767
BERT_BC	0.838	0.840	0.745	0.856	0.873	0.869	0.836
W_BC	0.846	0.844	0.749	0.865	0.870	0.872	0.858
W_FPT_BC	0.849	0.843	0.738	0.892	0.899	0.878	0.866
W_FPT_BC+pseudo labelling	**0.860**	0.843	0.758	0.910	0.895	0.881	0.871
W_FPT_BC+domain dictionary+pseudo labelling	0.850	0.844	0.762	0.929	0.913	0.881	0.873
RWL_BC+domain dictionary+ pseudo labelling	0.847	0.849	0.758	0.918	0.914	0.884	0.873
RWL+domain dictionary+pseudo labelling	0.854	**0.855**	**0.765**	**0.931**	**0.922**	**0.893**	**0.882**
Fusion model	0.872	0.859	0.792	0.940	0.923	0.891	0.887

Note: The terms of BERT_BC, W_BC, W_FPT_BC, RWL, and RWL_BC refer to the BERT_BiLSTM-CRF, WoBERT_BiLSTM-CRF, WoBERT_FurtherPretraing_BiLSTM-CRF, RoBERTa-wwm-large, and RoBERTa-wwm-large-BiLSTM-CRF model, respectively.

It is also noted from Table 2 that the use of pseudo labelling technique can help the WoBERT_BiLSTM-CRF model increase its strict *F*1 score by 0.5%. However, it makes the relaxed *F*1 score down by 0.1% if we compare it with the W_FPT_BC model in Table 3. This indicates that the pseudo labelling technique does not necessarily improve the performance of the medical NER models. One possible reason is that the generated pseudo labels may bring more noise to the original training data. For example, "前列腺" (prostate) is an entity of "anatomic part". It should be labelled as an "anatomic part" entity in the clause of "复发侵及前列腺" (recurrent invasion of prostate). However, "前列腺" (prostate) in the clause of "前列腺增生" (prostatic hyperplasia) is combined with other characters to form a "disease & diagnosis" entity. The pseudo labelling algorithm used in our paper somehow still labelled "前列腺" (prostate) in this clause as an "anatomic part" entity.

Table 3. NER model results (relaxed).

	Disease & diagnosis	Imaging examination	Laboratory examination	Surgery	Drug	Anatomic part	Total
BiLSTM-CRF	0.840	0.879	0.818	0.822	0.928	0.853	0.857
BERT_BC	0.923	0.889	0.805	0.956	0.950	0.927	0.908
W_BC	0.959	**0.901**	0.828	**0.976**	0.956	0.951	0.946
W_FPT_BC	0.980	0.893	0.810	0.970	0.961	0.961	0.955
W_FPT_BC+pseudo labelling	0.973	0.886	0.819	0.971	0.963	0.962	0.954
W_FPT_BC+domain dictionary+pseudo labelling	0.982	0.893	**0.829**	0.973	0.965	0.965	0.959
RWL_BC+domain dictionary+ pseudo labelling	0.976	0.889	0.820	0.967	**0.977**	0.960	0.956
RWL+domain dictionary+pseudo labelling	**0.987**	0.891	0.821	0.962	0.972	**0.968**	**0.961**
Fusion model	0.970	0.886	0.847	0.975	0.966	0.957	0.953

Note: The terms of BERT_BC, W_BC, W_FPT_BC, RWL, and RWL_BC refer to the BERT_BiLSTM-CRF, WoBERT_BiLSTM-CRF, WoBERT_FurtherPretraing_BiLSTM-CRF, RoBERTa-wwm-large, and RoBERTa-wwm-large-BiLSTM-CRF model, respectively.

For the results of each medical category, it is noted from Table 3 that the *F*1 scores of the W_FPT_BC model are 0.4%, 0.9%, and 0.6% higher than those of the RWL model for the category of imaging examination, laboratory examination, and surgery, respectively. This indicates that the W_FPT_BC model can recognize more short entities than the RWL model.

By observing the predicted results of the three models (that is, “WoBERT_Further-PreTrain-ing_BiLSTM-CRF” (W_FPT_BC), “RoBERTa-wwm-ext-large” (RWL), and “RoBERTa-wwm-ext-large_BiLSTM-CRF” (RWL_BC)), we found that the RWL model has a very good recognition rate on long entities, while the W_FPT_BC has a very good recognition rate on short entities. Taking the clause of “慢性肾衰, 尿毒症期” (chronic renal failure, uremic phase) as an example, we should recognize and label the entity in the clause as a “disease & diagnosis” entity. Among the three models, the RWL model can recognize the “disease & diagnosis” entity precisely, while the W_FPT_BC model only recognizes “慢性肾衰” (chronic renal failure) as a “disease & diagnosis” entity.

The overall recognition rate of the RWL_BC model is better than the other two models. For the results of total strict *F*1 score, it is noted from Table 2 that the *F*1 score of the RWL_BC model is 0.9% higher than the W_FPT_BC and RWL models, respectively. Based on this observation, we created a set of rules in the fusion model to merge the predicted results of the three models. For example, a rule could be “when the RWL and W_FPT_BC models identify an entity at a certain location in a text and the length of the identified entity is greater than 6, then we use the predicted result (or the identified entity) of the RWL model; otherwise, we use the predicted result of the W_FPT_BC model.” Another rule could be “if the RWL_BC model predicts a result in a certain position of a text and the other two models cannot, the predicted result of the RWL_BC model is then considered the final one for the NER process.”

After integrating or fusing the predicted results of the three models (that is, "W_FPT_BC + domain dictionary + pseudo labelling", "RWL_BC + domain dictionary + pseudo labelling", and "RWL + domain dictionary + pseudo labelling"), it is noted from the last line of Tables 2 and 3 (or the fusion model) that all strict *F*1 scores have increased in all the medical categories except the "anatomic part" category, but the relaxed *F*1 scores in many medical categories have decreased. This indicates that integrating the predicted results of the three models is most likely effective for the strict indicators, but not for the relaxed indicators. The reason could be that many "anatomic part" entities have nested structures, which thus cause boundary errors in the fusion model. For example, "右上肺支气管" (right upper lobe bronchus) is an "anatomic part" entity. However, "右上肺" (right upper lung) and "支气管" (bronchus) could be recognized as two separate entities after the predicted results of three models are integrated. The relaxed *F*1 indicator focuses on the number of recognized entities rather than entity boundaries. Therefore, the recognized entities, such as "右上肺" (right upper lung) and "支气管" (bronchus) are removed from the predicted results. As a result, relaxed *F*1 scores are decreased.

It is also noted from the last line of Tables 2 and 3 (the line of fusion model) that for the laboratory examination category, the strict and relaxed *F*1 scores are both significantly lower than the other categories. One possible reason is that some "laboratory examination" entities, such as "中性粒细胞百分比" (percentage of neutrophils) and "随机血糖" (random blood glucose) only appear in the test set but not in the training set. To alleviate this problem, we need a larger medical domain dictionary with more "laboratory examination" entities.

5. CONCLUSION AND FUTURE WORK

This paper presents a medical NER approach which incorporates a medical domain dictionary and pre-trained language models in recognizing medical named entities. Focusing on how to extract knowledge from un-labelled medical texts, we have performed the following tasks:

1) Medical dictionary. We constructed a medical domain entity dictionary and integrated it into pre-trained language models so that the recognition rate (measured by *F*1 score) of named entities has been significantly improved.
2) In-domain training. We fed un-labelled medical texts to pre-trained language models for in-domain training. Experimental results show that the proposed NER approach using the in-domain training technique can improve the performance of named entity recognition in medical communities.
3) Pseudo labelling. A pseudo labelling method is used to augment the CCKS-2020-CEMR data set by automatically annotating un-labelled texts and marking them as pseudo labels. The experimental results show that the use of the pseudo labelling technique can improve the NER performance from the view of strict *F*1 scores.

Medical named entity recognition using un-labelled texts and medical records is a challenging task. This research creates a way of using medical dictionary, in-domain pre-trained models, and pseudo labelling techniques to extract un-labelled medical texts. Future work will include 1) tasks to explore for other natural language processing (NLP) techniques which can produce more effective performance in dealing with un-labelled medical data for named entity recognition and 2) studies in identifying out-of-vocabulary entities efficiently.

ACKNOWLEDGEMENTS

This work is supported in part by the Guangdong Science and Technology grant (No. 2016A010101033) and the Hong Kong and Macao joint research and development grant with Wuyi University (No. 2019WGAH21).

AUTHOR CONTRIBUTIONS

C.J. Wen (klaywen15@163.com) initiated the proposed medical NER approach and conducted the experiments. C.J. Wen, T. Chen (chentao1999@gmail.com), and X.D. Jia (xudong.jia@csun.edu) contributed to the final version of the manuscript. C.J. Wen and Jiang Zhu (pacer0112@gmail.com) contributed to data preparation. T. Chen and X.D. Jia provided guidance to the project.

DATA AVAILABILITY STATEMENT

The data sets generated and/or analyzed during the current study are not publicly available due to the fact that the data sets are produced by medical expert consultants of Yidu Cloud based on their own experience. The publicly released version of the data sets needs the consent of all expert consultants, but they are available from the corresponding author on reasonable request.

REFERENCES

[1] Lei, J., et al.: A comprehensive study of named entity recognition in Chinese clinical text. Journal of the American Medical Informatics Association 21(5), 808–814 (2014)

[2] Wu, G., et al.: An attention-based BiLSTM-CRF model for Chinese clinic named entity recognition. IEEE Access 7, 113942–113949 (2019)

[3] Wang, S., Li, S., Chen, T.: Recognition of Chinese medicine named entity based on condition random field. Journal of Xiamen University (Natural Science) 48, 349–364 (2009)

[4] Wang, Y., Liu, Y., Yu, Z.: A preliminary work on symptom name recognition from free-text clinical records of traditional Chinese medicine using conditional random fields and reasonable features. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, pp. 223–230 (2012)

[5] Xu, Y., et al.: Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries. Journal of the American Medical Informatics Association 21(e1), e84–e92 (2014)

[6] Wu, Y., et al.: Named entity recognition in Chinese clinical text using deep neural network. Studies in Health Technology and Informatics 216, 624–628 (2015)

[7] Yang, J., et al.: Chinese electronic medical record named entity and entity relationship corpus construction. Journal of Software 27(11), 2725–2746 (2016)

[8] Yang, H., et al.: Named entity recognition based on bidirectional long short-term memory combined with case report form. Chinese Journal of Tissue Engineering Research 22(20), 3237–3242 (2018)

[9] Chowdhury, S., et al.: A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records. BMC Bioinformatics 19, 449 (2018)

[10] Wan, L., Luo, Y., Zhi, L.: The recognition of naming entity of Bi-LSTM Chinese electronic medical records based on the joint training of Chinese characters and words. China Digital Medicine 14(2), 54–56 (2019)

[11] Wu, Y., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)

[12] Lee, D.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Proceedings of ICML 2013 Workshop: Challenges in Representation Learning (WREPL), pp. 1–6 (2013)

[13] Liu, Y., et al.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

[14] Cui, Y., et al.: Pre-training with whole word masking for Chinese BERT. arXiv preprint arXiv:2004.13922 (2020)

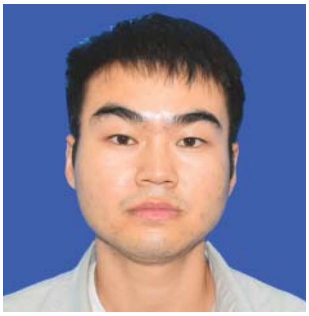
[15] Gururangan, S., et al.: Don't stop pretraining: Adapt language models to domains and tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8342–8360 (2020)

AUTHOR BIOGRAPHY

Chaojie Wen received his B.S. degree in Computer Science and Technology from Wuyi University in 2019. He is currently pursuing his M.S. degree in Electronic and Communication Engineering at Wuyi University, Guangdong, China. His research interests include natural language processing, named entity recognition, and relation extraction.
ORCID: 0000-0002-9325-9147

Tao Chen received his B.Eng. degree in Communication Engineering in 2003 from Nanjing Institute of Communication Engineering, China, and his M.Eng. degree in Computer Application in 2013 from Wuyi University, Guangdong, China, and his PhD degree in Computer Application from Harbin Institute of Technology, China in 2018. He joined Wuyi University, China, as a lecturer, in 2018. He is the author of one book, 17 articles and 3 patents. His research interests include natural language processing, deep learning, knowledge acquisition and reasoning, and sentiment analysis.
ORCID: 0000-0002-3634-0854

Xudong Jia is a Visiting Scholar of Wuyi University, China. He is also a Professor and the Associate Dean of College of Engineering and Computer Science, California State University, Northridge. He received his B.S. in 1983 and M.S. in 1986 from Beijing Jiaotong University, his M.S. in 1992 from University of Toronto, Canada, and his PhD in 1996 from Georgia Institute of Technology. His research interests include intelligent transportation systems (ITS) standards, geographic information system (GIS) applications in transportation, traffic safety, transportation information systems, travel demand management, and air quality. He is an associate editor of *IEEE Intelligent Transportation Systems Society* and *IEEE Open Journal of Intelligent Transportation Systems*.



Jiang Zhu received his B.S. degree in Electronic Information Science and Technology from Inner Mongolia University for Nationalities in 2018. He is currently pursuing his M.S. degree in Pattern Recognition and Intelligent System at Wuyi University, Guangdong, China. His research interests include natural language processing, named entity recognition, and data mining.